ABSTRACT

The statistical analysis described and claimed is a predictive statistical tree model that overcomes several problems observed in prior statistical models and regression analyses, while ensuring greater accuracy and predictive capabilities. Although the claimed use of the predictive statistical tree model described herein is directed to the prediction of a disease in individuals, the claimed model can be used for a variety of applications including the prediction of disease states, susceptibility of disease states or any other biological state of interest, as well as other applicable non-biological states of interest. This model first screens genes to reduce noise, applies k-means correlation-based clustering targeting a large number of clusters, and then uses singular value decompositions (SVD) to extract the single dominant factor (principal component) from each cluster. This generates a statistically significant number of cluster-derived singular factors, that we refer to as metagenes, that characterize multiple patterns of expression of the genes across samples. The strategy aims to extract multiple such patterns while reducing dimension and smoothing out gene-specific noise through the aggregation within clusters. Formal predictive analysis then uses these metagenes in a Bayesian classification tree analysis. This generates multiple recursive partitions of the sample into subgroups (the "leaves" of the classification tree), and associates Bayesian predictive probabilities of outcomes with each subgroup. Overall predictions for an individual sample are then generated by averaging predictions, with appropriate weights, across many such tree models. The model includes the use of iterative out-of-sample, cross-validation predictions leaving each sample out of the data set one at a time, refitting the model from the remaining samples and using it to predict the hold-out case. This rigorously tests the predictive value of a model and mirrors the real-world prognostic context where prediction of new cases as they arise is the major goal.